Latent Variable Analysis

Path Analysis Recap

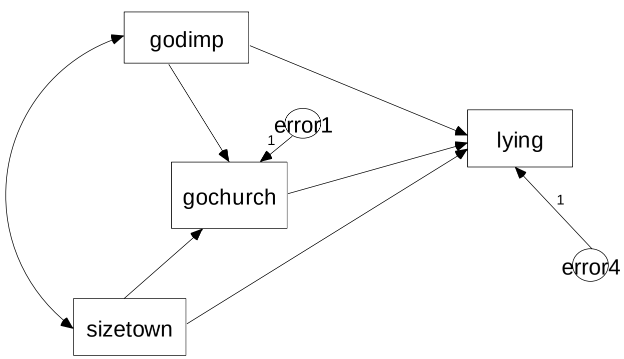
- **I. Path Diagram**
  - ***a. Exogeneous vs. Endogeneous Variables***
  - ***b. Dependent vs, Independent Variables***
  - ***c. Recursive vs. Non-Recursive Models***
- **II. Structural (Regression) Equations**
  - ***Normal Equations***
- **III. Estimating Path Coefficients**
- **IV. Identification**
  - ***a. Degrees of freedom***
  - ***b. Just Identified Models***
  - ***c. Overidentified Models***
  - ***d. Underidentified Models***

Path Analysis Recap

- **IV. Rules of decomposing the relationship between two variables**
- ***1. The components***
  - ***a. Direct effect***
    - path coefficient
  - ***Compound effects***
  - ***b. Indirect effect***
    - Start from the variable (Y) later in the causal chain to your right. Trace backwards (right to left) on arrows until you get to the other variable (X). You must always go against straight arrows (from arrow head to arrow tail).
  - ***c. Spurious effect (due to common causes)***
    - Start from variable Y. Trace backwards to a variable (Z) that has a direct or indirect effect on X. Move from Z to X.
  - ***d. Correlated (unanalyzed) effect***
    - It is like an indirect effect or a spurious effect due to common causes, except it includes one ,*and only a single one,* double headed arrow.
- ***2. Calculate compound paths by multiplying (path and/or correlation) coefficients encountered on the way***
  - ***Sewall Wright's rules***
  - **No loops**
    - Within one path you cannot go through the same variable twice.
  - **No going forward then backward**
    - Only common causes matter, common consequences (effects) don't.
  - **Maximum of one curved arrow per path**
- ***3. Add up all direct and compound effects***
  - **The sum is the total association**
    - In a just identified model the total association equals Pearson's correlation coefficient

Example: A just identified model

Determinants of honesty
Simple model with observed dependent and independent variables



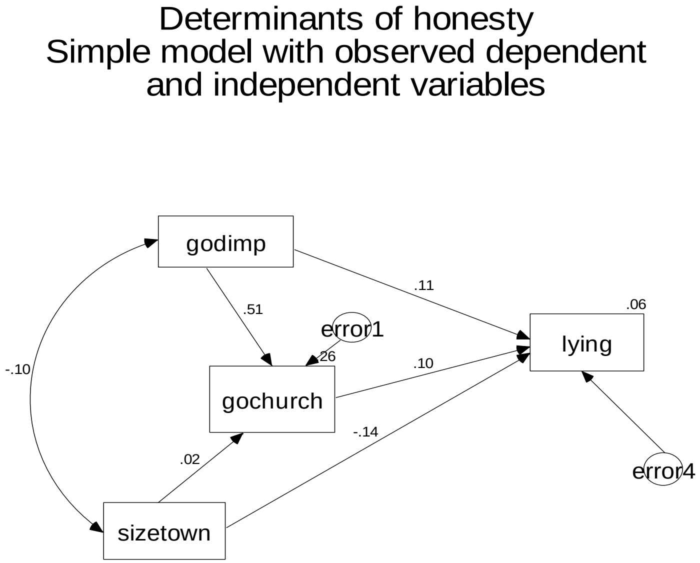
Correlations

		lying	sizetown	gochurch	godimp
lying	Pearson Correlation	1	-.158**	.160**	.175**
	Sig. (2-tailed)		.000	.000	.000
	N	1732	1732	1732	1732
sizetown	Pearson Correlation	-.158**	1	-.034	-.102**
	Sig. (2-tailed)	.000		.163	.000
	N	1732	1732	1732	1732
gochurch	Pearson Correlation	.160**	-.034	1	.508**
	Sig. (2-tailed)	.000	.163		.000
	N	1732	1732	1732	1732
godimp	Pearson Correlation	.175**	-.102**	.508**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	1732	1732	1732	1732

**. Correlation is significant at the 0.01 level (2-tailed).

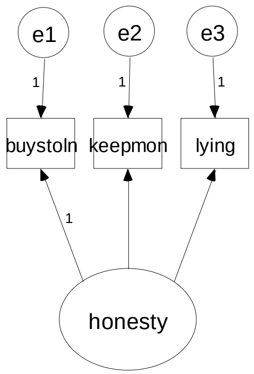
6 equations (correlations)
6 unknowns (5 paths and 1 correlation)

Standardized Estimates



Latent Variable and Its Indicators

Estimating the latent variable separately



3 equations (correlations)
3 unknowns (paths)

Correlations

		buystoln	keepmon	lying
buystoln	Pearson Correlation	1	.276**	.371**
	Sig. (2-tailed)		.000	.000
	N	1732	1732	1732
keepmon	Pearson Correlation	.276**	1	.457**
	Sig. (2-tailed)	.000		.000
	N	1732	1732	1732
lying	Pearson Correlation	.371**	.457**	1
	Sig. (2-tailed)	.000	.000	
	N	1732	1732	1732

**. Correlation is significant at the 0.01 level (2-tailed).

The three observed variables are indicators of the latent variable Honesty which is a concept. They are effect indicators because they are the effects of the latent variable.

Structural Equations:

(1) $B=p_{bh}*H+e1$
(2) $K=p_{kh}*H+e2$
(3) $L=p_{lh}*H+e3$

Normal Equations:

If we just multiply each equation by its independent variable we will not get anywhere. Take the 1st equation:

$r_{bh}= p_{bh} *r_{hh}+r_{he1}$ $r_{hh}=1$ and $r_{he1}=0$ so $r_{bh}= p_{bh}$ but what is r_{bh}?

So we must multiply each equation by the other two

(1) $B=p_{bh}*H+e1$ multiplied by (2) $K=p_{kh}*H+e2$

$B*K=(p_{bh}*H+e1)*(p_{kh}*H+e2)= p_{bh}*H*p_{kh+} *H+p_{bh}*H*e2 + p_{kh}*H*e1+ e1*e2$

Turn it into a normal equation

$r_{bk} = p_{bh}*p_{kh}* r_{hh} +p_{bh}*r_{he2}*+ p_{kh}*r_{he1} +r_{e1e2}$

because $r_{hh} =1$ and $r_{he2} =0$ and $r_{he1} =0$ and $r_{e1e2} =0$

$\mathbf{r_{bk} = p_{bh}*p_{kh}}$ this also follows from the rules of decomposing relationship between two variables K and B are related only through their common cause of H

the same way we can calculate two other normal equations:

$\mathbf{r_{bl} = p_{bh}*p_{lh}}$

$\mathbf{r_{lk} = p_{lh}*p_{kh}}$

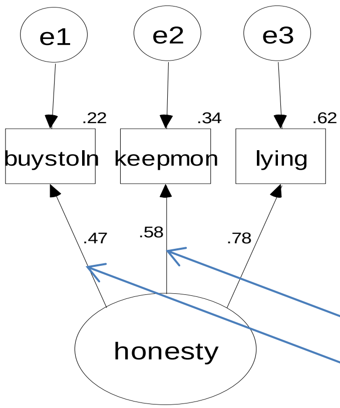
Finding the Path Coefficients

- **Normal Equations:**

- **(1) $r_{bk} = p_{bh}*p_{kh}$**
- **(2) $r_{bl} = p_{bh}*p_{lh}$**
- **(3) $r_{lk} = p_{lh}*p_{kh}$**
- We express p_{bh} from (1)
- $r_{bk} / p_{kh} = p_{bh}$
- We substitute p_{bh} in (2)
- $r_{bl} =(r_{bk} / p_{kh})*p_{lh}$
- **We express p_{lh}**
- $r_{bl} /(r_{bk} / p_{kh})=p_{lh}$
- **We substitute p_{lh} in (3)**
- $r_{lk} =(r_{bl} /(r_{bk} / p_{kh})) *p_{kh} = p_{kh} * p_{kh} * r_{bl} / r_{bk}$ → $p_{kh}^2= r_{lk} * r_{bk} / r_{bl}$
- p_{kh}^2= .457*.276/.371 = .34 → $p_{kh} = \sqrt{.34}$ **<u>=+/-.583</u>** *Notice that this number can be +.583 or -.583 because the latent*
- *variable can be scaled in either direction (it can measure honesty or dishonesty).*
- *We choose +.583 and the latent variable will be scaled in the same direction as K.*
- We can get p_{bh} by substituting in (1)
- .274=p_{bh} *.583 → **<u>p_{bh} =.470</u>**
- And we can get p_{lh} by substituting in (3)
- .457= p_{lh} *.583 → **<u>p_{lh} = .784</u>**

The Measurement Model Calculated by STATA

Estimating the latent variable separately



Correlations

		buystoln	keepmon	lying
buystoln	Pearson Correlation	1	.276**	.371**
	Sig. (2-tailed)		.000	.000
	N	1732	1732	1732
keepmon	Pearson Correlation	.276**	1	.457**
	Sig. (2-tailed)	.000		.000
	N	1732	1732	1732
lying	Pearson Correlation	.371**	.457**	1
	Sig. (2-tailed)	.000	.000	
	N	1732	1732	1732

**. Correlation is significant at the 0.01 level (2-tailed).

$r_{bk}=p_{hb}*p_{hk}=.47*.58\approx.276$

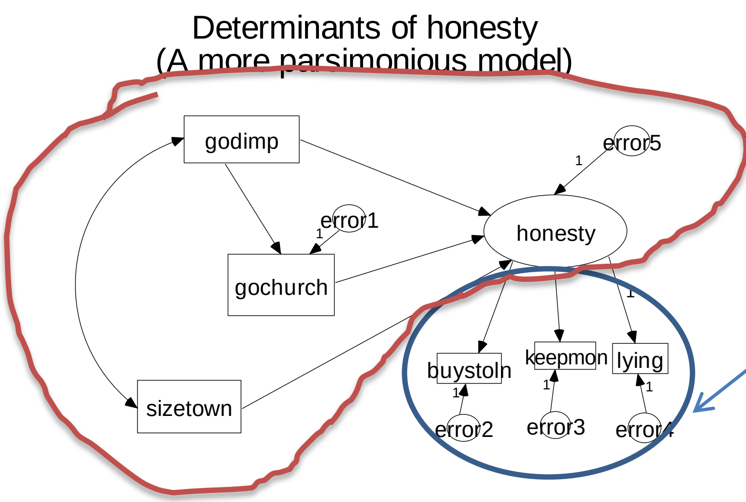
$r_{bl}=p_{hb}*p_{hl}=.47*.78\approx.371$

$r_{lk}=p_{hl}*p_{hk}=.78*.58\approx.457$

The paths and R-squareds tell us how good each indicator is measuring the latent variable.

Attitude about lying (LYING) is the best indicator of honesty (.78). 62 percent of what people say about their attitude about lying reflects their attitude about honesty. The rest is error (e3).

Causal Model with Latent Variable

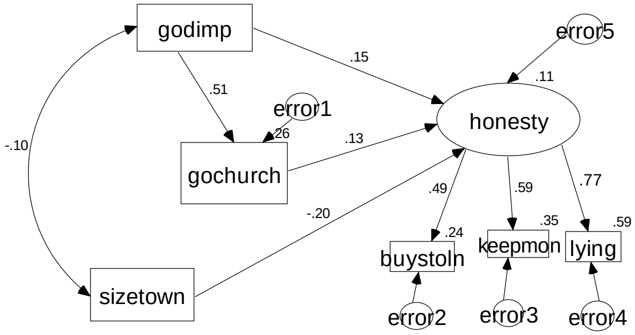


- Notice that we have 7 paths and 1 correlation or 8 coefficients to estimate.
- We have 6*(6-1)/2=15 normal equations (correlations)
- We have 15-8=7 degrees of freedom
  - We can test the entire model
- The model has a
- substantive part (relationships among concepts) and a
- measurement part (relationships among concepts and indicators).

- *IMPORTANT:*
- *Measurement CANNOT be separated from substantive theory. In fact, STATA estimates the two simultaneously. If you change the substantive model, the measurement model may change as well.*

Evaluating Your Output

Determinants of honesty
(A more parsimonious model)



- Things to look for:
- 1. ***Could STATA do the job?***
  - Did the model converge?
    - It should have no error message AT THE LAST STEP like
      - non-concave function encountered
      - unproductive step attempted
- 2. ***Is your measurement model good?***
  - Are the indicators strong enough?
    - Direct effects of latent variables on indicators
  - Are their relative weights reasonable?
- 3. ***What does your substantive model say?***
  - Direct effects path coefficients
  - Indirect effects
- 3. ***How well are you predicting endogenous variables?***
  - Fitting each endogenous variable
    - R-squared
- 4. ***Did you draw the right model/picture?***
  - Fitting the ***entire*** model
  - Chi-squared test – statistical significance
    - Does the model significantly diverge from the data?
  - Various fit measures
    - How much does the model diverge on some standardized scale

How STATA Fits Your Model

Sample Correlations

	sizetown	godimp	gochurch	lying	buystoln	keepmon
sizetown	1.000					
godimp	-.102	1.000				
gochurch	-.034	.508	1.000			
lying	-.158	.175	.160	1.000		
buystoln	-.129	.158	.108	.371	1.000	
keepmon	-.130	.128	.125	.457	.276	1.000

Fitted Correlations

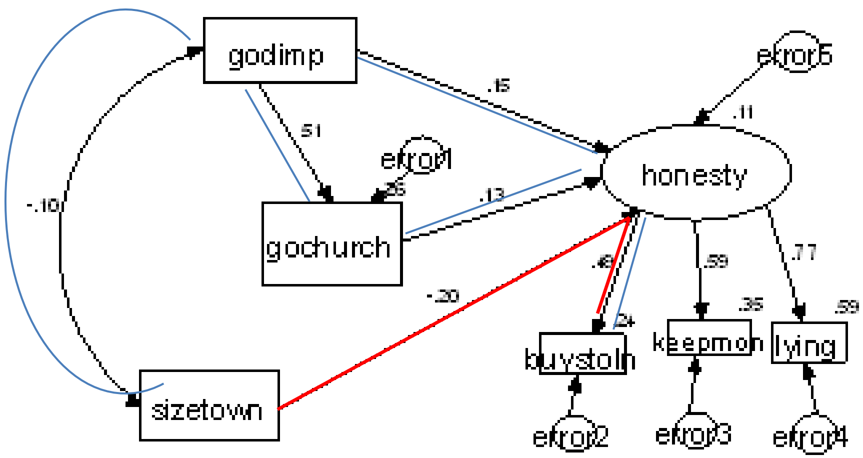
	sizetown	godimp	gochurch	lying	buystoln	keepmon
sizetown	1.000					
godimp	-.102	1.000				
gochurch	-.052	.508	1.000			
lying	-.168	.183	.164	1.000		
buystoln	-.107	.116	.104	.373	1.000	
keepmon	-.130	.141	.127	.452	.288	1.000

The fit of the entire model is evaluated by comparing the observed and implied correlations (covariances). *(STATA really works with unstandardized variables and uses covariances rather than correlations. But for the sake of simplicity we assume that the world is standardized.)*

STATA compares these two tables as you did in 205 when you calculated Chi-squared for a table comparing cell by cell the predicted (or implied) and the observed values. There you compared frequencies, here STATA compares correlations (covariances).
Notice that here your model is good if Chi-squared is NOT significant because it means that the discrepancy between your model's predictions and the data is insignificant.
Also notice that

Let's take the correlation between BUYSTOLN and SIZETOWN.
Observed: -.129,
Implied: -.107.
Our model does not predict this correlation very well.
How is the implied correlation computed?
It is computed using the rules of path analysis.

The Implied Correlation Between BUYSTOLN and SIZETOWN



No direct effect

Indirect effect through HONESTY

-.20*.49=-.098

No spurious effect due to common causes (SIZETOWN is exogenous)

Correlated/Unanalyzed effects

through GODIMP and HONESTY

-.10*.15*.49= -.007

through GODIMP and GOCHURCH and HONESTY

-.10*.51*.13*.49=-.003

Implied correlation is (-.098)+(-.007)+(-.003)=-.108 ≈-.107

Evaluating the Fit of the Entire Model

LR test of model vs. saturated: chi2(7) = 8.73, Prob > chi2 = 0.2725

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(7)	8.731	model vs. saturated
p > chi2	0.273	
chi2_bs(14)	1349.373	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.012	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	0.033	
pclose	1.000	Probability RMSEA <= 0.05
Information criteria		
AIC	44387.168	Akaike's information criterion
BIC	44496.309	Bayesian information criterion
Baseline comparison		
CFI	0.999	Comparative fit index
TLI	0.997	Tucker-Lewis index
Size of residuals		
SRMR	0.011	Standardized root mean squared residual
CD	0.306	Coefficient of determination

- Chi-squared (chi2):
  - Measure of *statistical significance* of the fit (it is like the F-statistics for R-squared)
  - A Chi-squared is big if
    - You have a poor fit and/or you have a large N
  - Here our Chi-squared is 8.726 with 7 degrees of freedom
  - The probability level tells you the likelihood of getting this discrepancy between implied and observed correlation/covariance by chance when in the population your model would have a perfect fit (0.2725)
    - Your Chi-squared is NOT significant at the .05 or .1 level. It means that your fit is GOOD. The discrepancy is insignificant.
- Measures of FIT
  - It measures how close the path coefficients reproduce the correlation/covariance matrix (it is like R-squared)
    - model – your model
    - Saturated model – model with 0 degree of freedom (d.f.)
    - Baseline --- all paths (but not correlations) are set to 0
  - **RMSEA**: the fit close if the lower bound of the 90% CI is below 0.05 and label the fit poor if the upper bound is above 0.10
  - Akaike information criterion (**AIC**) and Bayesian (or Schwarz) information criterion (**BIC**) are used not to judge fit in absolute terms but instead to compare the fit of different models. Smaller values indicate a better fit.
  - Comparative Fit Index (**CFI)** and Tucker–Lewis Index (**TLI)**, two indices such that a value close to 1 indicates a good fit. TLI is also known as the nonnormed fit index.
  - A perfect fit corresponds to a Standardized Root Mean Squared (SRMR) of 0. A good fit is a small value, considered by some to be limited to 0.08.
  - Coefficient of Determination (**CD**) a perfect fit corresponds to a value of 1 and is like a R-squared for the whole model.